Semantic Video Content Indexing and Retrieval using Conceptual Graphs

Mbarek Charhad, Georges Quénot

Semantic Video Content Indexing and Retrieval using Conceptual Graphs

Mbarek Charhad & Georges Quénot
CLIPS-IMAG, BP, 38041
Grenoble Cedex 9
{Mabrek.Charhad, Georges.Quenot}@imag.fr

Abstract

In this article, we propose a conceptual model for video content description. This model is an extension of the EMIR² model proposed for image representation and retrieval. The proposed extensions include the addition of some views such as temporal and event views that are specific to video documents, the extension of the structural view to the temporal structure of video documents, and the extension of the perceptive view to motion descriptors. We have kept the formalism of conceptual graphs for the representation of the semantic content. The various concepts and relations involved can be taken from general and/or domain specific ontologies and completed by lists of instances (individuals). The proposed model has been applied on TREC video 2002 and 2003 corpora that mainly contain TV news and commercials videos.

Keywords: Multimedia information retrieval, conceptual indexing, video document, ontology

1. Introduction

Advances in multimedia technologies have made possible the storage of huge collections of video documents on computer systems. In order to allow an efficient exploitation of these collections, it is necessary to design tools for content-based access to their documents. As this is the case for text documents, keyword based indexing and retrieval can be used (from speech transcript and/or closed captions for instance). Concept based indexing is an improvement over keyword based indexing because it removes the ambiguities between keyword senses due to synonymy and polysemy. Conceptual graph [1] based indexing is even better since not only non-ambiguous concepts are used but also relations between these concepts are indexed.

In the case of video, there are a number of specificities due to its multimedia structure. For instance, a given concept (person, object ...) can be present in different ways: it can be seen, it can be heard, it can be talked of, and combination of these representations can occur. Of course, these distinctions are important for the user. A query involving X as "Show me a picture of X" or as "I want to know what Y has said about X." are likely to give quite different answers. The first one would look for X in the image track while the second would look in the audio track for a segment in which Y is the speaker and X is mentioned in the transcription. In addition, among all possible relations that could be represented in conceptual graphs, some are especially appropriate for content-based video indexing. Here, we propose a model for the indexing and retrieval of semantic video content using conceptual graphs. This model is an extension of the EMIR² (1) model for the indexing and retrieval of image semantic content using conceptual graphs [2].

Video structure should be analogous to text document structure, where we perform a structural analysis to decompose it into sections, paragraphs, sentences, and words. Similarly, to facilitate fast and accurate content access to video data, we should segment a video document into shots and scenes, compose a table of contents, and extract key frames or key sequences as index entries for scenes or stories. Therefore, the core research in content-based video retrieval concerns the development of methods for automatically parsing video, audio, and text in order to identify meaningful composition structures and for extracting and representing content attributes for any video source.

This paper is organised as follow. In section 2, we present some approaches and tools for modelling video by content. The section 3 deals with our proposition for model video content using conceptual graphs and an ontology structure to enrich video description. In section 4, we detail the use of our system for the indexing of video by semantic content. We conclude this paper by listing some futures works and perspectives.

2. Previous work

A number of researchers have proposed techniques to model video content. Some of these techniques are based on describing physical objects and spatial relationships between these objects. An approach that uses spatial

(1) EMIR²: Extended Model for Image Representation and Retrieval.

relationships for representing video semantics is proposed in [9]. In this approach, images are processed and represented by spatio-temporal logic. The prototype provides a novel query interface by which query-by-sketch is employed to query video contents. However, due to the limitations of the methodology used in this approach, the modelling of higher-level concepts, like spatio-temporal events, is not addressed. Moreover, no method for the ranking of retrieved video data has been proposed. The framework discussed in [8] defines a set of algebraic operators to allow spatio-temporal modelling.

Other techniques for modelling of video content depend on the semantic classification of video content. For example, the model proposed in [3] allows hierarchical abstraction of video expressions representing scenes and events which provides indexing and content-based retrieval mechanisms. It allows users to assign multiple interpretations to a given video segment and provides functionalities for creating video presentations. An object hierarchy is built using 'is-a' generalizations, and interval inclusion based inheritance is introduced to capture the hierarchical flow example we can automatically describe the video generated by a monitoring camera. Indeed, in this kind of video, the content is known and moving objects are of an expected type (cars). Since all these models are purely semantic-based, they have the inherent limitations of being annotation-based and application-dependent that make them restricted to specific domain.

Petkovic and Jonker propose in [5] a model making it possible to model the events in video sequences and consider the four following information's layers (low-level with the high-level): pixels, characteristics, objects, and events. Thus, it is possible to detect specific events by defining the states and the interactions between suitable objects.

In the context of multimedia information retrieval, we distinguish two types of video system indexing and retrieval. There exist the systems known as generic [4], which makes it possible to obtain a classification of the various video sequences available without taking into account information of contextual nature. These systems allow, for instance, classifying the various video sequences according to the type of scene (indoors, outdoors) and the camera motion (static or moving). On the other hand, there are systems known as specific that allow indexing of only a particular kind of video, such as for instance TV news, or sports events such as the soccer game. In this case, indexing is known as contextual [6]. The specific systems, even if their use is limited to a specific type of video sequence nonetheless make it possible to answer many queries from users of video indexing and retrieval systems. They depend strongly on the context they are dedicated to and generally limited to. Indeed, indexing of the video sequences can be seen like a classification of sequences of images, each sequence possibly corresponding to an event of expected type. The problem returns to detection of such events.

3. A model for conceptual video content description

To enable search and retrieval of video within large archives, we need a good description of video content. In this work, we consider that video input is segmented into shots. Most visual and audio features (motion, speech, text) will be used to associate a description to each shot. For example, in order to describe the content of video news, we use some concepts to describe scenes like meeting, speech, interview, live reporting or events/topics like sports, politics and commercials. We use also the identities of persons that can be recovered form the visual flow (person who appears on the screen), form audio or from textual information.

We propose to develop the description attributed to such shots by using relationships between concepts. These relations can be spatial (on the right, on the left…), to describe positions, and/or temporal, for modelling events sequences (before, after…). We also use semantic relations to describe actions (speaks, meets, etc.) within a formalism of knowledge representation.

In this work, we use conceptual graphs (GCs) [7] [1] which constitutes formalism for knowledge representation. They express meaning in a form that is both logically precise and humanly readable. With their graphic representation, conceptual graphs serve as a readable, but formal design and specification language. Conceptual graphs have been implemented in a variety of projects for information retrieval, database design, expert systems, and natural language processing. The following figure shows a simple description that is represented by conceptual graphs:

Example: "A person is between a rock and a hard place"

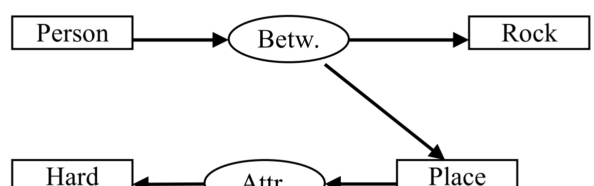

Figure 1 : Example of representation using GCs

The same example may be represented in the following text form:

[Person] -> (Betw) -> [Rock]

-> [Place] -> (Attr) -> [Hard].

3.2 Structural and semantic description of video content

The description of video content from the viewpoint of its composition is structured around segments that represent physical spatial, temporal and/or spatio-temporal [14] components of the video content. Each segment may be described by signal-based features (color, texture, shape, motion, and/or audio features) and some basic semantic information. Physical segmentation of video is based on shot detection [?]. Shot is defined like continuations of images resulting from a continuous camera acquisition. The shot is often considered as the smallest temporal unity for a video sequence (however, sub-shot segmentation is also considered). We use the word "scene" to define a unit of several successive shots representing the same domain.

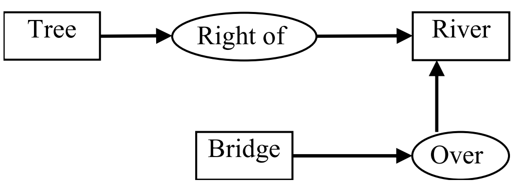

Figure 2: A shot description with spatial relationships

The analysis of video contents indicates that the main relationships can be grouped into three types: spatial, temporal and action.

- ✓ **Spatial relations:** this type of relationships describes the positions of two visual objects in video content. Examples are 'in front of', 'on', 'under', etc. To describe video content using spatial parameter, we mainly use visual information but this does not exclude to find also this information in audio flow (speech) containing a spatial description of content. Figure 2 shows an example of modelling using spatial relationships.
- ✓ **Temporal relations:** the temporal aspect is specific to video documents. Relationships of this type make it possible to order or synchronize events. The known relationships used for the temporal aspect are Allen's temporal relationships [10] (starts, meets, after…). In our model, we exploit the temporal segmentation to infer semantic description based on temporal aspect such start time and end time of each shot. This kind of information is important to synchronise events for example.
- ✓ **Action relations:** Relationships which describe an action are typically related to the description of events occurring in videos. The action relationships can be further divided into two types: the mutual and the directional. In mutual relationships, the related entities are symmetrical. For example: "a man is shaking hands with a boy" is the same as "a boy is shaking hands with a man". In directional relationship, the types of related entities need to be differentiated.

3.1 Extended model for video description

As we already mentioned in the abstract, our model is an extension of the EMIR² model for the indexing and retrieval of still images using conceptual graphs [2].

The EMIR² model involves four types of relations between concepts that are well suited for still image content description. These correspond to four different views of image content:

- ➢ A structural view, involving relations between parts and subparts,
- ➢ A spatial view, involving 2D (within the image) or 3D (in the real world represented in the image) relations between image parts,
- ➢ A symbolic view, involving relations that define symbolic properties of image parts,
- ➢ A perceptive view, involving relations that define perceptive properties (color, texture …) of image parts.

The proposed video content model extends the EMIR² model. For individual images or for segments whose visual content is stable, the EMIR² views and relations apply directly. For other aspects of videos, the following views and relations are extended or added:

- An extended structural view, involving relations between parts and subparts but with parts considered as temporal segments,
- A temporal view, involving temporal relations between video segments;
- An extended perceptive view, involving relations that define additional perceptive properties (mainly about motion) of video parts;
- An event-based view, involving relations that define what happens to or between video parts.

3.3 Video content description ontology

Due to the fact that different users may have different interpretations of a same video segment and due to word synonymy and polysemy, annotators may use different instances of concepts to describe a same audiovisual element.

In order to help solving or reducing the problem of mapping annotation element to high level concepts, we propose to conceive specific domain ontologies containing concepts to be used to annotate video segments. A list of terms called instances comes with each concept.

Figure 3 represents a part of a knowledge base used to help matching video high-level description and video annotation elements.

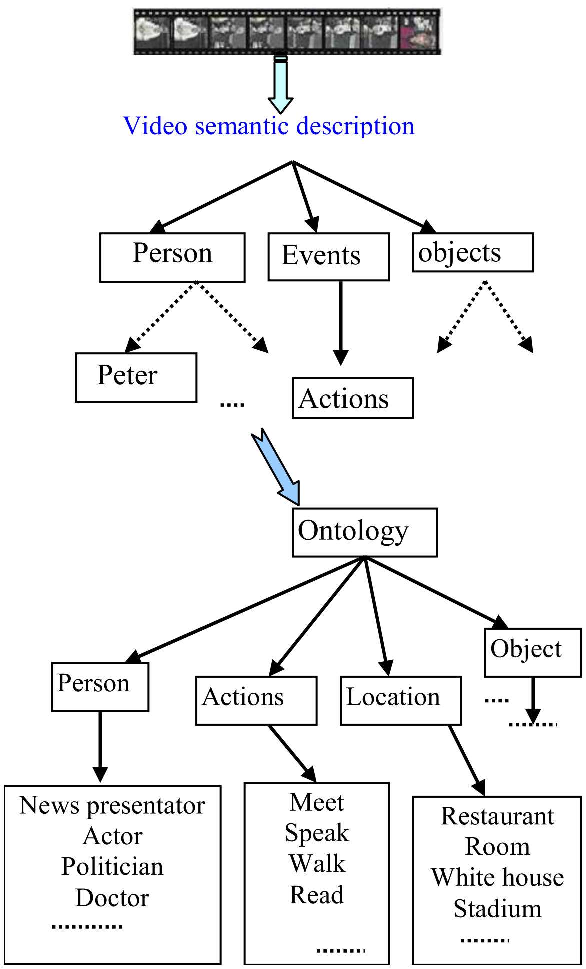

Figure 3 : Ontology for video content description

“Ontology” is a specification of an abstract, simplified view of the world we wish to represent for some purpose. Therefore, ontology defines a set of representational terms that we call *concepts.* Inter-relationships among these concepts describe a target world. Ontology can be constructed in two ways, domain specific and generic. WordNet is an example of generic ontologies.

The “Ontology” term used to describe and represent a field of knowledge [11]. “Ontology” is used in applications that need to share a domain of information/knowledge. A domain is a field of knowledge of specific interest such as medicine. Ontologies are expressed in a logical language so that detailed, precise, coherent, of good direction, and explicit distinctions can be made classes, properties, and relations. Ontology-based applications can automatically process information by using ontology structures, and thus provide evolved services to intelligent applications like conceptual / semantic research.

Ontologies can appear very useful as means of structuring descriptions of video semantic content. Ontologies could support semantic descriptors for libraries of images, sounds, or other objects. Ontology-based annotations can be used for indexing and research processes. Being understood that several people can describe these non-textual objects in various manners, it is significant that the functionalities of research exceed the use of simple keywords like research topics. Ontology for non-textual objects has the following functionalities:

- It provides taxonomy of terms. This taxonomy can be used to generalize or restrict research topics. A search system is able to use the ontology structure for query expansion for instance.
- It should express and support the representation of default knowledge. This functionality is essential for a real exploitation of the semantics of a query.

4. Experimentation

Experimentations have been conducted using part of the TREC video 2002 and 2003 corpora. We have used automatic transcription of speech to extract terms that may be considered as a concept or instance of concept. From the visual flow, we have generated a maximum of semantic relationships (walking, speaking, meeting…) from a manual collaborative annotation provided in the context of the Video-Annex system [13]. We have enriched all the annotations of the corpora by using relational description. Indeed, to each concept we have associated a semantic relationship that may be a spatial description, temporal or an activity.

Figure 4 illustrates an example of a schema that may be used as an input for our video search system to retrieve video segments from a semantic description. Examples of query associated to this description for may be a punctual graph such as we which find all video segments that contain “*Kofi Annan speaks with Clinton*”. (Figure 4, (b)).

Like in the EMIR² system, the correspondence between the query and any document is searched as a projection of the graph representing the query on the graph representing the document. We also use ontology to match concepts that form the query and the video description.

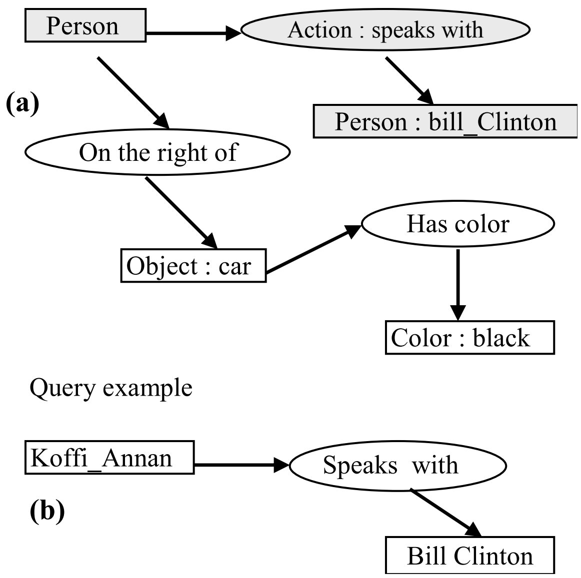

Figure 4 : An example of content modelling using CGs.

Operation of projection is based on the relation of specification / generalization, in our example, the graph (b) is a specification of the graph (a). Indeed, there exists in the graph (a) a sub graph which represents a generalization of query graph. The graph in greyed represents the projection of (b) in (a).

Figure below shows the current interface of the developed video retrieval system.

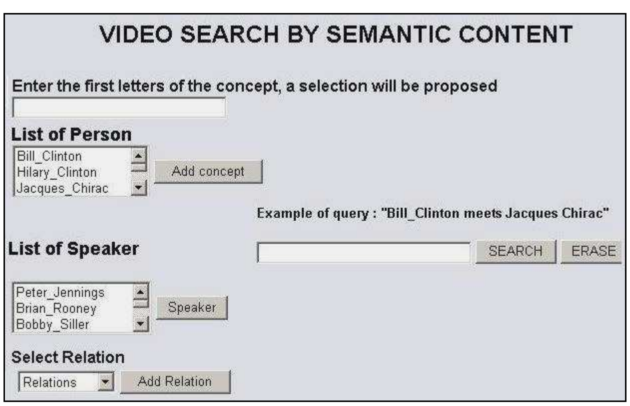

Figure 5: Video search by semantic content using conceptual graphs

5. Conclusion

In this article, we have proposed a conceptual model for video content description. This model is an extension of the EMIR² model proposed for image representation and retrieval. The proposed extensions of include the addition of some views such as temporal and event views that are specific to video documents, the extension of the structural view to the temporal structure of video documents, and the extension of the perceptive view to motion descriptors. We have kept the formalism of conceptual graphs for the representation of the semantic content.

The proposed model allows video content representation both at the semantic and at the signal levels. The signal based representation is done via the perceptive view using classical signal-level descriptors. The representation by conceptual graph allows a more precise description of the content and it also allows the linking between the involved concepts and the sub-medias (audio, image and text) of the video documents.

The various concepts and relations involved in the model can be taken from general and/or domain specific ontologies and completed by lists of instances (individuals for instance). We have presented in section 3 of this paper a schema of an ontology that allows resolving linguistic problems such as synonymy or polysemy when describing video content.

Finally, our model is applied on TREC video 2002 and 2003 corpora that mainly contain TV news and commercials videos. This category of video may be considered as a structured document when we analyse the content such as reportages or studio settings. Our work consisted for a large part in modelling the content using concepts and conceptual relationships.

The next step of this work is, first, to extend the modelling scheme with more concepts and conceptual relationships and, second, to test our system on other categories of video like documentaries, sports or movies

6. References

[1] Sowa John F, “Conceptual Structures: Information Processing in Mind and Machines”, Addison-Wesley publishing company, 1984.

[2] Mourad Mechkour, “EMIR²: An Extended Model for Image Representation and Retrieval”, in DEXA'95 Database and Expert system Applications, London pp 395-404 September, 1995.

[3] Xingquan Zhu, Jianping Fan, Ahmed K. Elmagarmid, Xindong Wu, “Hierarchical video content description and summarization using unified semantic

and visual similarity", Multimedia Systems, v.9 n.1, p.31-53, July 2003.

[4] Giuseppe Amato, Gianni Mainetto, Pasquale Savino: "An Approach to a Content Based Retrieval of Multimedia Data", Multimedia Tools and Applications, 1998.

[5] M. Petkovic and W. Jonker, "A framework for video modelling", in IASTED International Conference on Applied Informatics, Innsbruck, Autriche, Février 2000.

[6] Silvia Hollfelder, André Everts and Ulrich Thiel, "Designing for Semantic Access: A Video Browsing System", Multimedia Tools and Applications 11(3): 281-293, 2000.

[7] Nastaran Fatemi and Philippe Mulhem, "A Conceptual Graph Approach for Video Data Representation and Retrieval", Third International Symposium, IDA-99, Amst, the Netherlands, August 1999.

[8] F. Golshani and N. Dimitrova, "Retrieval and delivery of information", in multimedia database systems Information and Software Technology, 36(4):235-242, May 1994.

[9] S. Beretti, Alberto Del Bimbo, Enrico Vicario: Efficient Matching and Indexing of Graph Models in Content-Based Retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10): 1089-1105, 2001.

[10] Rune Hjelsvold, Roger Midtstraum, "Modelling and Querying Video Data", proceedings of the 20 th VLDB Conference Santiago, chile, 1994.

[11] Noy, N.F. and McGuinness, "Ontology Development: A Guide to Creating Your First Ontology", Available as SMI technical report SMI-2001-0880, 2001.

[12] Dechilly T. and Bachimont B, " Une ontologie pour éditer des schémas de description audiovisuels, extension pour l'inférence sur les descriptions", Actes de IC'00, Toulouse, 2000.

[13] Ching-Yung Lin, Belle L Tseng and John R Smith, "VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning" IEEE Intl. Conf. on Multimedia & Expo (ICME), Baltimore, July 2003.

[14] Di Zhong and Shih-Fu Chang "Spatio-temporal Video Search Using the Object Based Video Representation", Published in the International Conference on Image Processing, (ICIP'97), in Santa Barbara, CA, 1997.